Noname manuscript No.
(will be inserted by the editor)

Duration modelling and evaluation for Arabic statistical parametric speech synthesis

Imene Zangar · Zied Mnasri · Vincent Colotte · Denis Jouvet

the date of receipt and acceptance should be inserted later

Abstract Sound duration is responsible for rhythm and speech rate. Furthermore, in some languages phoneme length is an important phonetic and prosodic factor. For example, in Arabic, gemination and vowel quantity are two important characteristics of the language. Therefore, accurate duration modelling is crucial for Arabic TTS systems. This paper is interested in improving the modelling of phone duration for Arabic statistical parametric speech synthesis using DNN-based models. In fact, since a few years, DNN have been frequently used for parametric speech synthesis, instead of HMM. Therefore, several variants of DNN-based duration models for Arabic are investigated. The novelty consists in training a specific DNN model for each class of sounds, i.e. short vowels, long vowels, simple consonants and geminated consonants. The main idea behind this choice is the improvement that we already achieved in the quality of Arabic parametric speech synthesis by the introduction of two specific features of Arabic, i.e. gemination and vowel quantity into the standard HTS feature set. Both objective and subjective evaluations show that using a specific model for each class of sounds leads to a more accurate modelling of the phone duration in Arabic parametric speech synthesis, outperforming the state-of-the-art duration modelling systems.

Imene Zangar
University Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis, Electrical Engineering Department, BP 37, 1002 Tunis, Tunisia, E-mail: imene.zangar@enit.utm.tn

Zied Mnasri
University Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis, Electrical Engineering Department, BP 37, 1002 Tunis, Tunisia, E-mail: zied.mnasri@enit.utm.tn

Vincent Colotte
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France,
E-mail: vincent.colotte@loria.fr

Denis Jouvet
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France,
E-mail: denis.jouvet@loria.fr

Keywords Parametric Arabic speech synthesis · duration modelling · Deep neural network · Objective and subjective evaluation

1 Introduction

Text-To-Speech synthesis (TTS) has been historically used as help for people with visual impairments, for reading ebooks and messages. Now it is being used worldwide in a large range of applications varying from consumer electronics, talking devices and especially as a cost-effective alternative to human speakers, in charge of broadcasting routine messages like news bulletins, weather forecasts and traffic alerts. More recently, it started to be also used for language learning and online translation.

Overview of Text-To-Speech systems

Several TTS techniques have been developed since more than half a century: formant synthesis [1], articulatory synthesis [2], concatenative methods, i.e. diphone synthesis [3,4] and unit selection synthesis [5], and parametric speech synthesis based either on hidden Markov models (HMM) or on deep neural networks (DNN). In fact, this timeline order depends on the progress of storage abilities and of computational resources. Hence, the first methods, i.e. formant synthesis and diphone synthesis need only a small database, whereas unit selection synthesis has been feasible only when it had been possible to store huge databases of speech segments. Finally, statistical parametric synthesis methods, based on HMM and DNN, were promoted by the development of complex algorithms able to deal with the corresponding computational load. Therefore, each technique has its advantages and drawbacks that depend on the background approach: whereas formant synthesis and diphone synthesis do not require high memory, the output sound quality is not quite satisfactory. On the opposite side, unit selection synthesis provides a high quality sound, at the cost of using a huge database. Finally, parametric speech synthesis represents a trade-off between relatively small memory and high quality of sound.

Parametric speech synthesis

The first parametric speech synthesis technique was based on HMM to model and predict the speech parameters, i.e. duration, fundamental frequency (F0) and spectrum. Decision trees were used to share context-dependent parameters, and thus to select the best-fitting HMM parameters according to the available contextual information [6]. The generated parameters were fed into a vocoder, like MLSA (Mel Log Spectrum Approximation) [7], to generate the output speech signal. Therefore, it was possible to develop this technique using less speech data than for unit selection. However, the naturalness of its output speech has been so far less appreciated than that of unit selection synthesis,

mainly due to oversmoothing, acoustic modelling and vocoding [8]. DNN-based speech synthesis is another parametric speech synthesis approach. Instead of relying on HMMs and associated decision trees to predict the speech parameters, DNNs map the input contextual features to the corresponding speech parameters [9]. Then, in the same way, the output speech is generated by a vocoder. Furthermore, recent relevant developments in sequence-to-sequence DNN-based TTS such as Tacotron [10,11] have given good subjective evaluation results. These models outperform state-of-the-art results through combining alignment, duration and acoustic parameters modelling. It should be noted that in the recent benchmark evaluations of TTS techniques, parametric speech synthesis, either based on HMM or DNN, has always been ranked amongst the best quality techniques, notwithstanding the development of novel techniques, mainly based on deep learning [56]. Such a good rating can be explained by the particularity of parametric speech synthesis, based on modelling prosody specifically for each target language.

Prosody modelling

Prosody modelling is a key component in any TTS system, especially in parametric speech synthesis. Prosody refers to intonation, rhythm and loudness [12]. Phonetically speaking, these aspects are expressed by F0, duration and energy. Except unit-selection TTS, where signal modification is avoided [5], most concatenative speech synthesis systems rely on signal modification to adapt the values of the prosodic parameters of the database units to the predicted ones, in order to ensure intelligibility and naturalness of the output speech. In the case of parametric speech synthesis, duration and acoustic parameters (F0 and spectrum) are modelled either by HMM or by neural networks. These models rely on label features that cover most of contextual, linguistic and phonological information that may influence the output parameters, i.e. duration, F0 and spectrum. A medium size database, e.g. 450 sentences in [6], was used to train the context-dependent HMM models, nevertheless larger databases lead to better speech synthesis quality. The context-dependent HMMs are clustered using a decision tree, whose binary questions refer to the label features. The binary tree is constructed by iteratively selecting the most relevant question, that is the question that leads to the largest log-likelihood gain of the training data. The size of the decision tree depends on some threshold of the training criterion, i.e. MDL (Minimum Description Length) value [13] or cross-validation [6].

Prosody modelling for parametric TTS

The main advantage of parametric speech synthesis consists in its ability to model the acoustic parameters for a wide range of contexts. In fact, thanks to decision trees clustering where parameters distributions are shared, the effect

of unseen contexts and data sparsity is reduced. Though parametric speech synthesis has been proved to produce accurate prosody models in many languages, such as Japanese [6], English [8], Arabic [14], etc., it did not succeed to reach the unit-selection speech quality as reported in [15]. This is due to several reasons, namely vocoding, acoustic modelling and oversmoothing. As far as vocoding is concerned, since the output speech is generated using MLSA filter [7], based on a pulse train and white noise excitation, a buzz effect is usually perceived. To cope with this issue, many vocoders were tried out, to enhance the quality of the output speech, such as STRAIGHT [15] and more recently WORLD [16]. For acoustic modelling, though decision trees are able to match the contextual features to the adequate HMM model, the whole system suffers from the simple frame-wise acoustic parameter modelling [9]. Actually, speech and more particularly prosody parameters generation is a continuous process, which might be considered as recurrent rather than simply overlapping. To resolve this problem, many modifications were brought to the HMM structure, to enhance their ability to model acoustic parameters taking into account their inherent dependencies [9]. The last issue, i.e. oversmoothing, is due to using statistical averaging in the training phase, to reduce the data sparsity. Besides, in the synthesis phase, dynamic constraints are used. Both techniques yield in over-smoothed trajectories of the acoustic parameters (F0 and spectrum), so that a muffling effect is perceived by the listener, because the micro-prosodic variations cannot be captured [9]. To overcome this effect, simple solutions like post-filtering or explicitly using the training data as constraints in the parameter prediction algorithm were suggested [9]. Another solution is to use multiple-level statistics, like global variance (GV) when generating speech parameters trajectories [9].

Duration modelling for parametric TTS

Several approaches have been developed to model segmental duration since the beginning of TTS technology. These approaches can be classified into two main families: rule-based approaches and data-driven ones. In the first family, an explicit formulation of segment duration is found out, where phone and/or syllable duration is calculated using analytic formulas [1], or assuming the existence of a common compression/extension factor for all phone durations within a syllable [17]. In [18], the duration of a segment is calculated as the sum of products of a hypothetic intrinsic duration multiplied by some contextual factors like phoneme class, stress and place of articulation. However, such empiric rules do not provide high accuracy, and were criticized for the lack of fundamental and consistent proofs [12]. In the second family, segment duration is obtained through training. Particularly, in HMM-based speech synthesis, state duration distributions are estimated. Each state duration is explicitly modelled by a Gaussian distribution. Context-dependent decision trees are used to cluster the duration distributions. In the synthesis part, the state duration is obtained by mapping the contextual feature vector to the corre-

sponding state duration HMM. Finally, the phone duration is obtained by summing its state durations [6]. Nevertheless, the main critic of HMM-based speech synthesis, and in particular HMM-based phone duration modelling is the averaged output, since HMM models are mostly relying on Gaussian distribution [8]. Therefore, a better way to model the micro-variations of prosody, including phone duration, is looked for through replacing HMM by DNN in parametric speech synthesis.

Arabic parametric TTS

Modern Standard Arabic (MSA), which is widely used among all Arab-speaking countries as the official and literary language, has 28 consonants and three vowels, /a/, /u/ and /i/ [19]. Most consonants could be geminated (doubled) which is indicated in writing through adding the specific diacritic sign (shadda) whereas each vowel has a short and a long version. One of the main components of prosody is segmental duration. In fact, rhythm, speech rate, and partly intonation, depend on units durations. Moreover, in some languages like Arabic, a change in vowel quantity (long vs. short) and in gemination alters the meaning of the word, e.g. (عرض) /aradha/ "he presented", (عارض) /**a:**radha/ (he objected) and (عرّض) /a**rr**adha/(he exposed).

Application of TTS to the Arabic language has started since the beginning of TTS technology. Unit selection TTS was successfully adapted to Arabic [20] as well as the HMM-based TTS toolkit (HTS) [14]. Also in [14], the synthesis filter was modified to improve its quality: the set of acoustic parameters consisted of spectrum amplitude and multi-band voicing decision, and the MLSA filter has been replaced by a multi-band excitation vocoder [21]. Prosody and more particularly duration modelling has been studied specifically for Arabic to be integrated into Arabic TTS using diphone synthesis techniques like PSOLA and MBROLA. Such duration models were based either on rule-based methods, i.e. explicit formulas for phoneme duration in [22], or data-driven techniques like artificial neural networks (ANN) in [23, 24]. However, prosody modelling for Arabic can be improved by implementing specific models, taking care of its characteristics. In fact, such an approach has been proved to be efficient in improving the overall quality of Arabic speech synthesis, in our previous work [25].

Paper's scope and organization

Since the prosodic module is one of the key components of any parametric speech synthesis system, the work presented in this paper aims to improve prosody modelling in general, and phone duration modelling in particular. This goal is met by finding out DNN-based duration models that allow enhancing the quality of HMM-based parametric speech synthesis for Arabic. DNN-based duration models are developed using additional linguistic features of Arabic, i.e. vowel quantity and gemination, which have recently been successfully introduced into HTS system for Arabic TTS [25]. The novelty of the proposed approach consists in establishing a specific duration model for each class of Arabic phonemes. The duration predicted externally by the specific DNN models is then introduced into the HTS system for generating F0 and spectrum features, and then the speech output. Both objective and subjective evaluations show that using a specific model for each class of sounds leads to a more accurate modelling of the phone duration. The objective evaluation results have already been briefly presented in [26], whereas in this paper it will be extended and a subjective assessment will be added.

The rest of the paper is organized as follows, Section 2 presents duration modelling for parametric speech synthesis, starting from the baseline HMM model to current DNN ones, including other external models tried out for parametric speech synthesis. Section 3 describes DNN and recurrent neural network (RNN) duration models that we investigated for Arabic, with a special care to recall the modifications of the input features set to meet Arabic language requirements, that have been recently proposed for Arabic parametric speech synthesis using HMM [25]. Section 4 details the objective evaluation measures and subjective listening test results. Finally, a discussion and a conclusion end the paper.

2 Duration modelling for parametric speech synthesis

In HMM-based parametric speech synthesis, prosody and acoustic modules rely on HMMs to generate duration and excitation features, i.e. log(F0) and the spectral parameters (and their dynamic counterparts Δ and $\Delta\Delta$). However, in order to improve naturalness, other approaches have been investigated in a bid to replace the HMM duration model.

Data-driven duration modelling for parametric speech synthesis

HSMM duration model for parametric speech synthesis

For the modelling of the state durations in parametric speech synthesis systems, explicit state duration models have been introduced. As it is conventional in context-dependent HMMs, context-dependent parameters are clustered through the use of decision trees. This is also the case for the state

duration distributions; and the decision trees are built using the contextual features (previous and following phonemes, position in the syllable, in the word, in the phrase, etc.). During synthesis, the duration Gaussian distributions used are those returned by the decision tree, according to the contextual features, and the state durations are determined in order to maximise the duration probabilities.

Hidden-semi Markov models [27,28] based on explicit duration Gaussian models were used to model duration in the HMM-based speech synthesis system (HTS) [29]. However, they did not considerably improve the naturalness of parametric speech synthesis output speech in comparison to unit selection synthesis, as reported in [8].

Other data-driven duration models for parametric speech synthesis

One of the most perceived weaknesses of parametric speech synthesis consists is bland prosody, partly due to over-averaging, particularly in duration, which results from the use of Gaussian distributions [30]. To enhance the ability of parametric speech synthesis to generate more natural rhythm and speech rate, many duration models were proposed to refine or to replace the original HSMM one. For the refinement methods, the HSMM-based state duration modelling was extended to the phoneme and/or the syllable level in [31,32]. In the same way, the diagonal Gaussian distribution of the HSMM model was replaced by a full-covariance Gaussian distribution in [33], or by a gamma distribution, with fine tuning of the gamma parameter [34]. The second option consists in simply replacing the HSMM duration model by a better performing one. Several data-driven duration modelling techniques have shown high performance in modelling segmental duration for diphone synthesis, such as artificial neural networks (ANN) [17], classification and regression trees (CART) [35], support vector regression (SVR) [36] and multi-adaptive regression splines (MARS) [37]. Therefore, many machine learning based models were externally trained to provide more accurate state, phone or syllable durations for parametric speech synthesis [38].

State duration modelling using MLP instead of decision trees

In [38], external SVR and Multi-layer perceptron (MLP) models explicitly provide state or phone duration to the parametric speech synthesis system. Another approach consists in replacing the context-based decision tree of duration HMMs by an MLP, in a hybrid HMM-MLP scheme [39]. This is done in a two-stage process, where the state durations (in number of frames) are obtained by training monophone HMMs and Viterbi state alignment. In the second stage, the obtained state durations are trained as outputs of the MLP, whose inputs are the contextual, linguistic and phonological features, which used to be fed into the decision tree. Objective evaluation of phone duration prediction has shown a decrease in RMSE (ms) while measured for the overall phonemes, and partial decrease with respect to phoneme identity, mainly for

vowels and semi-vowels. Listening tests have shown rather a preference for the hybrid model, though the "no preference" answer has been dominating.

DNN-based duration modelling for parametric speech synthesis

Thanks to their proved ability to approximate continuous functions, DNN have become the state-of-the-art tool in regression problems. Particularly, duration modelling for speech synthesis has been achieved using several approaches with DNN.

DNN-based state duration distribution vs. segmental duration models

DNN-based duration modelling for parametric speech synthesis can be used in two ways. In the first case, DNN are used instead of decision trees in the HSMM duration model to map the label features to the pre-trained state duration HMMs; whereas in the second case, raw segmental durations, i.e. state, phone or syllable durations, are directly estimated from the label features using DNN.

- *DNN-based state duration distribution modelling*
  Since the work of [9], DNN have been preferred to decision trees to model state duration HMMs for parametric speech synthesis due to some limitations of decision trees. Actually, decision trees are not well-fitted to model complicated functions like XOR, d-bit parity function or multiplex problems [9]. Furthermore, decision trees have a poor generalization power, since they process input data by partitioning the input space into regions, each associated with a terminal node. This yields in neglecting some "weak" features [40]; whereas using DNN involves training all the input data while updating weights, which means a better generalization. On the other side, the outputs are the parameters of duration HMM, i.e. the mean value and the variance. Actually, predicting these statistic parameters and then generating durations using the Gaussian distribution ensure average values. It should be also mentioned that in this internal scheme, only state durations can be predicted, like in HSMM-based duration modelling, then the phone duration is obtained as the sum of the durations of its states.
- *DNN-based segmental duration modelling*
  In this model, raw segmental duration, i.e. state, phone or syllable duration is directly predicted from the label features using DNN. In the synthesis stage, the predicted durations are included to the parametric speech synthesis system. However, the way phone durations are predicted may depend on the DNN architecture. In some works, like [41], discrete phone durations are directly predicted from label features using DNN, whereas in [42] phone durations were predicted by adding an explicit constraint, which consists in the mean square error between the phone duration and the sum of the

phone's state durations (L_{ec}) (cf. (1)):

$$L_{ec} = \sum_{j=1}^{N} (\sum_{s=1}^{5} \hat{d}_{j,s} - d_j)^2, \quad (1)$$

where N is the number of phonemes, $\hat{d}_{j,s}$ and d_j are respectively the predicted duration of state s within phone j and the target duration of phone j.

DNN architectures for duration modelling

Duration modelling may also take advantage from the different architectures of DNN, i.e. feedforward, recurrent or multi-task learning, which have all been already applied for this purpose.

- *Feedforward-DNN*
  A feedforward-DNN is the simplest DNN architecture. Based on MLP structure, current networks usually contain more hidden layers than former MLP. Actually, feedforward-DNN's benefit from the advances of GPU to include many hidden layers, which might increase the prediction accuracy [43]. Feedforward DNN can be used for classification, if output parameters are a finite set of discrete labels, or for regression, in case of continuous values. Since duration is a regression task, with continuous values, feedforward-DNN are well fitted to model it. Therefore, feedforward-DNN models have been used, internally as a replacement of context-based decision trees to predict state duration distributions, as in [9], or externally to estimate raw duration.
- *LSTM neural networks*
  In feedforward-DNN, features are processed without recurrence. This means that the sequential aspect of speech, in particular affecting rhythm and speech rate, is neglected. However, this aspect is primordial in speech production. Recurrent neural networks (RNN) are based on forward or backward time propagation of the input features. Amongst RNN variants, long-short-term memory (LSTM) have been particularly praised for addressing the vanishing gradient problem in standard RNN's [44]. In LSTM, the output at each layer is generated from its input and a memory cell, which is obtained from the previous time memory cell and a set of gates (input, output and forget gates) in addition to the current layer's input itself [43]. Hence, at each layer, each feature passes through the input, forget and output gates in addition to the memory cell, before the output is obtained. This ensures that the previous phonemes are taken into account while calculating the current phoneme duration.
- *BLSTM neural networks*
  Furthermore, forward and backward information propagation in LSTM can be combined. Such variant is called bi-directional LSTM (BLSTM).

In this scheme, the hidden activation functions are defined as positive-direction or negative-direction, respectively for forward or backward direction. The same for the weight matrices and the recurrent matrices. LSTM and BLSTM were both successfully utilized to model state durations in [45, 46] and raw phone duration in [41,47,48] for parametric speech synthesis.

Continuous vs. discrete duration modelling using DNN

In the case of continuous phone duration modelling, the optimization criterion is the mean square error L_{mse_i} expressed for each utterance i having n_i phonemes (cf. (2)):

$$L_{mse_i} = \frac{1}{n_i}\sum_{j=1}^{n_i}(d_{i,j} - \hat{d}_{i,j})^2, \tag{2}$$

where $d_{i,j}$ and $\hat{d}_{i,j}$ are the reference and the predicted phone durations respectively of the phoneme sequence $(v_{i,1}, v_{i,2}, ..., v_{i,n_i})$ of the i^{th} utterance [41].

Another strategy consists in the estimation of the discrete probability distribution of the phone duration. The interest of this method lies in estimating the probability of the current phone duration given the phoneme sequence history $(v_{i,1}, v_{i,2}, ..., v_{i,j})$ to be expressed by the cross-entropy criterion L_{ce_i} for each utterance u_i (cf. (3)):

$$L_{ce_i} = -log(p(\hat{d}_i|v_i)) = -\sum_{j=1}^{n_i} log(p(\hat{d}_{i,j}|v_{i,1}, ..., v_{i,j})), \tag{3}$$

where $p(\hat{d}_{i,j}|v_{i,1}, ..., v_{i,j})$ is the duration probability distribution of phoneme $v_{i,j}$ given the sequence $v_{i,1}, ..., v_{i,j}$ [41]. Though both criteria, i.e. L_{mse} for continuous duration and L_{ce} for discrete duration probability density are equivalent from the perspective of probability distribution [41], we have relied in this work on continuous duration modelling, since we use continuous values, instead of discrete probability densities of phone duration.

3 Proposed DNN-based duration modelling for Arabic parametric speech synthesis

Based on the aforementioned approaches, in particular those applied for DNN-based parametric speech synthesis, and aiming at taking care of Arabic speech requirements, we introduce a set of DNN-based models, specifically designed for Arabic parametric speech synthesis, along with the speech materials and the feature set.

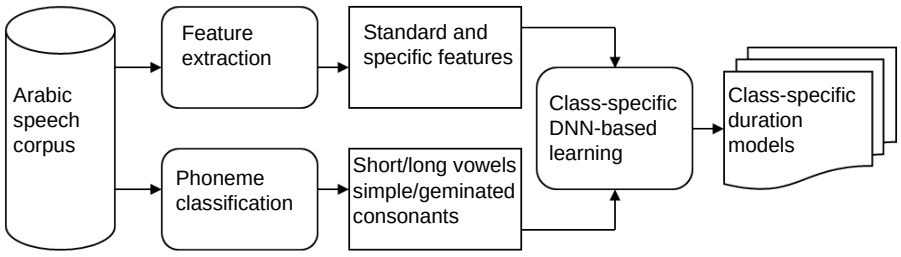

Fig. 1 Workflow of the proposed class-specific duration modelling approach: Each class of Arabic phonemes is processed independently to train its own DNN-based duration model.

Workflow of the proposed approach

The main idea behind the proposed approach consists in designing a dedicated DNN-based duration model for Arabic, and to introduce it into the standard HTS parametric speech synthesis toolkit [53]. As illustrated in Figure 1, the novelty consists in (a) adding some novel features to the standard HTS feature set [8], namely vowel quantity and gemination, (b) training a specific DNN duration model for each class of Arabic phonemes, namely short vowels, long vowels, simple consonants and geminated consonants, instead of using one duration model for all phonemes. The ultimate goal, and the real added value of this work consists in exploiting the results of the proposed DNN-based duration model as an external input to the HTS system, instead of the internal HMM-based duration model of HTS, in order to improve the performance of HTS, specifically for Arabic (cf. Figure 2).

Introduction of DNN-based duration model into HMM-based parametric speech synthesis system

Since our aim is to improve the quality of parametric speech synthesis for Arabic, using characteristic features of Arabic, i.e. gemination and vowel quantity, and since it has been proven that the aforementioned DNN models outperform all the state-of-the-art duration prediction techniques, we propose to improve the duration prediction performance of HMM-based parametric speech synthesis for Arabic, by using external DNN models, dedicated for duration prediction, instead of relying on the classical duration HMM's used in HTS system. This is feasible since HTS allows imposing parameters values, whether for excitation, i.e $log(F_0)$, spectrum or duration, so that the values provided by the internal HMM models will be replaced by those introduced from external sources. Figure 2 shows the whole HTS system, where only the duration HMM module is replaced by an external DNN. In the following, the steps of the experimental protocol will be detailed.

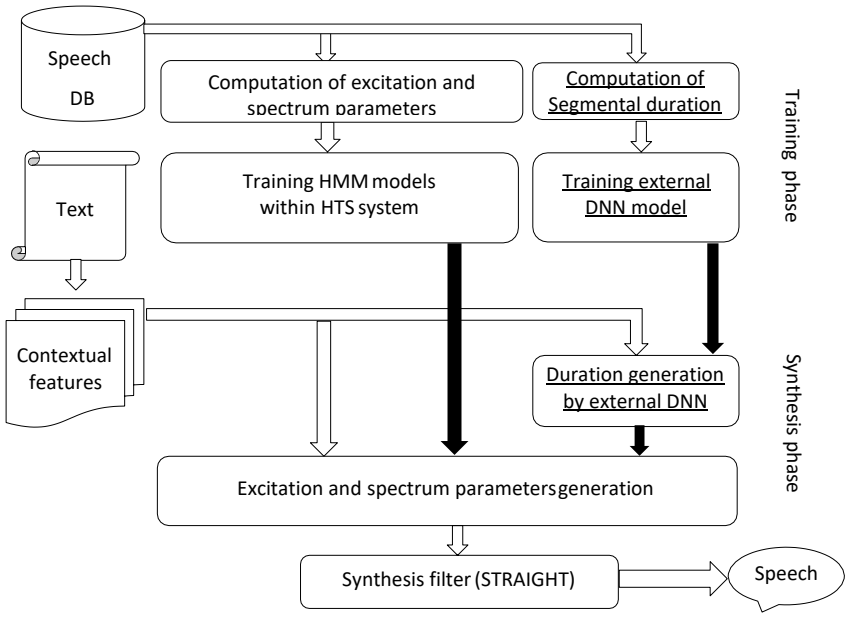

Fig. 2 Overview of the modified HMM-based parametric speech synthesis system with external duration DNN modelling

Speech material

In order to train the DNN-based duration models, a set of 1597 utterances was used. The utterances were taken from an Arabic speech corpus [55, 52], representing news bulletin in MSA (Modern Standard Arabic), read by a native-Arabic male speaker. The signals were recorded at 48 KHz sampling rate, with 16-bit precision. The corpus was divided into three subsets, approximately 70% of utterances for training, 20% for development and 10% for test. The label features and output target duration have undertaken the required preprocessing, i.e. label features coding and output durations normalization. Categorical label features were coded in binary, e.g. like stressed/not stressed syllables, or in discrete values, e.g. phoneme identity, whereas unlimited-value features like phoneme's position in the syllable were coded numerically. The label features were coded into a 445-coefficient vector. On the other side, output duration distribution was analysed and a log-transform was applied to normalize it. In fact, duration distribution normalization was suggested in [50] to increase the prediction ability of neural networks. In addition, [51] suggests that using an adequate normalizing transformation of input data may dramatically reduce the estimation errors and the calculating time during the training phase. Therefore the numeric contextual features were normalized in the interval $[0, 1]$ by dividing their values on the maximum value in the training set. Since the main contribution of this work concerns the modelling of phone duration by phoneme class, the reference durations have been statistically analysed by

phoneme class. Table 1 reports, for each class of sounds, the number of phone occurrences, the mean duration, and its associated standard deviation value.

Table 1 Statistics of phone durations by phoneme class for the training, validation and test sets

Dataset	Class	Number of occurences	Mean value (ms)	Standard deviation (ms)
Training	Simple consonant	37872	91	37
	Geminated consonant	4040	180	43
	Short vowel	23670	71	36
	Long vowel	11565	120	49
	Pauses	2458	340	156
Validation	Simple consonant	5850	95	43
	Geminated consonant	831	202	38
	Short vowel	3825	85	44
	Long vowel	1929	116	59
	Pauses	1132	445	211
Test	Simple consonant	3658	93	38
	Geminated consonant	450	202	46
	Short vowel	2466	82	38
	Long vowel	1004	138	61
	Pauses	548	454	126

Features extraction

The baseline contextual feature set is similar to the one used in Arabic speech synthesis using HTS [14], where the features can be classified into three main subsets:

- Positional features, e.g. position of the current phoneme in the current syllable (forward and backward), number of phonemes in the current syllable, etc.
- Linguistic features, e.g. identity of the current phoneme, guess-part-of-speech (content-word/non content-word), etc.
- Phonological features, e.g. lexical stress of the current syllable, etc.

Besides, two Arabic-language-specific features, i.e. gemination and vowel quantity were proved to enhance the quality of parametric speech synthesis for MSA [25]. Therefore, both were added to the feature set. The original duration were provided by automatic segmentation, using forced alignment [49]. In addition, a portion of the corpus was segmented manually by experts, to evaluate the quality of the automatic segmentation, and to verify whether the corpus needs a totally manual segmentation [49].

Benchmarking and proposed duration models

This work aims to improve the prediction of the phoneme durations for parametric synthesis, using DNN taking into account some characteristic features of Arabic (simple consonants/geminated consonants and short vowels/long vowels). Five duration models were trained and evaluated, all using the same set of contextual features and the same output reference durations:

1. HMM baseline model (from HTS toolkit [53]),
2. DNN model (from MERLIN toolkit [54]),
3. Feedforward-DNN model,
4. DNN-LSTM model,
5. DNN-BLSTM model.

Besides, feedforward-DNN, DNN-LSTM and DNN-BLSTM models were trained with various training subsets corresponding to

- either all the phonemes present in the training set,
- or only short vowels for the short vowel dedicated model, only long vowels for the long vowels dedicated model, only simple consonants for the simple consonant dedicated model, only geminated consonants for the geminated consonant dedicated model, and only pauses for the pause dedicated model.

Since learning performance is mainly adjusted empirically, several structures of the proposed models were tried out by varying the number of hidden layers, the number of their nodes, and the activation functions. As the task is to predict continuous values of phone duration, the RMSE minimisation criterion was adopted for training. The RMSprop optimizer was adopted in all the experiments. To avoid over-learning, early stopping was used. Thus, if L_{mse} evaluated on the development set does not improve after a certain number of epochs, set to 20 in our case, the training process is stopped. Table 2 shows, for each class of sounds, the DNN structure leading to the best duration prediction results on the development set.

The class-specific model is defined as the model which gives the best accuracy for the given class of sounds, independently of its performance on the other classes of sounds.

4 Evaluation protocol

A twofold evaluation process was carried out to assess the quality of the duration prediction, through an objective evaluation by measuring the differences between the reference durations and the predicted durations, and a subjective evaluation based on listening tests.

Objective evaluation

The objective evaluation process consists in comparing the performance of the class-specific DNN modelling to state-of-the-art models, i.e. HMM model as

Table 2 Description of the model structure leading to the best accuracy on the development set for each phoneme class and for pauses

Class of sounds	Model	Training set	Model description	# of epochs
simple consonants	DNN-BLSTM(1)	simple consonants	2 feedforward layers with 512 units each and activation function *tanh*, plus 2 BLSTM layers with 128 units each	94
geminated consonants	DNN-BLSTM(2)	geminated consonants	2 feedforward layers with 16 units each and activation function *tanh*, plus 2 BLSTM layers with 16 units each	74
short vowels	DNN-BLSTM(1)	short vowels	2 feedforward layers with 512 units each and activation function *tanh*, plus 2 BLSTM layers with 128 units each	89
long vowels	Feedforward-DNN	long vowels	2 feedforward layers with 512 and 256 units resp. and activation function *tanh*	97
pauses	DNN-LSTM	all the phonemes	3 LSTM layers with 1024, 512 and 512 units resp.	95

used in HTS, a standard DNN model trained on all phonemes, DNN model as used in MERLIN, and the former MLP model developed for Arabic [24]. To select the best class-specific DNN models, many architectures were tried out. Then, they are selected as the best models on the development set (c.f. Table 2) because they give the best scores and finally they were kept to be evaluated on the test set. The standard DNN model trained on all phonemes is built using three LSTM layers having 1024, 512 and 512 nodes respectively. The DNN model as used in MERLIN is composed by 6 hidden feedforward layers with 1024 units each and tangent hyperbolic (*tanh*) as activation function. This model relies on the same set of features as HTS. The MLP model from [24], contains two hidden layers with 26 units each, and uses *sigmoid* and *tanh* as activation functions. This model does not use the same set of linguistic features as HTS. Evaluation is carried out by focusing on the comparison of the performance of the models on each class of phonemes and on pauses; a global comparison is also reported that takes into account all the sounds.

Table 3 presents the evaluation of the prediction of the phone durations on the test set data, as measured by the following criteria: root mean square error (RMSE), mean absolute error (MAE) and correlation coefficient (corr) between original and predicted duration, as calculated respectively by (4a),

Table 3 Comparison of RMSE, MAE and correlation between predicted duration and reference duration on the test set for each phoneme class and for the various modelling approaches

Class of sound	Duration model	RMSE (ms)	MAE (ms)	Corr
simple consonants	HMM from HTS	**25**	18	0.76
	Class-specific DNN	**25**	**17**	**0.77**
	DNN-all-phone	28	20	0.72
	DNN-MERLIN	26	18	0.75
	MLP from [24]	35	25	0.50
geminated consonants	HMM from HTS	43	**31**	0.43
	Class-specific DNN	**42**	32	**0.51**
	DNN-all-phone	48	37	0.34
	DNN from MERLIN	54	40	0.15
	MLP from [24]	62	50	0.42
short vowels	HMM from HTS	**22**	**16**	0.82
	Class-specific DNN	**22**	**16**	**0.84**
	DNN-all-phone	23	17	0.82
	DNN from MERLIN	26	19	0.81
	MLP from [24]	26	19	0.78
long vowels	HMM from HTS	49	34	0.68
	Class-specific DNN	**40**	**28**	**0.77**
	DNN-all-phone	48	35	0.65
	DNN from MERLIN	54	38	0.66
	MLP from [24]	68	52	0.07
pauses	HMM from HTS	**109**	73	0.54
	Class-specific DNN	**109**	**70**	0.54
	DNN-all-phone	**109**	**70**	0.54
	DNN from MERLIN	146	110	**0.60**
	MLP from [24]	188	158	0.56

(4b) and (4c):

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(d_j - \hat{d}_j)^2}, \tag{4a}$$

$$MAE = \frac{1}{N}\sum_{j=1}^{N}|d_j - \hat{d}_j|, \tag{4b}$$

$$corr = \frac{cov(d, \hat{d})}{\sigma_d \sigma_{\hat{d}}}, \tag{4c}$$

where $cov(d, \hat{d})$ is the covariance of duration vectors d and $\hat{d}$; and σ_d, $\sigma_{\hat{d}}$ are the standard deviations of d and $\hat{d}$ respectively.

Table 3 shows that the proposed class-specific DNN models, lead to the best performance amongst all the evaluated models, i.e. baseline HMM model (from HTS), the DNN model trained on all phonemes (DNN-all-phone), the baseline DNN model from MERLIN and the former MLP model, for almost

every measure, RMSE, MAE and correlation coefficient, and almost every class of sounds, i.e. short/long vowels, simple/geminated consonants, and pauses. Table 4 presents the accuracy values computed on all the phonemes, as well as the accuracy computed on all sounds, i.e., on all the phonemes plus pauses; it shows that using class specific models improves the prediction duration accuracy, in comparison to state-of-the-art models. Figure 3 displays the original and predicted duration distributions of phone durations for each class of sounds on the test set. A good match is observed for simple consonants, for short vowels and for long vowels. However, for the geminated consonants, the predicted duration distribution is sharper, showing a lower standard deviation with higher mean values, which means that the predicted duration values are more concentrated around the average duration than are the reference duration values.

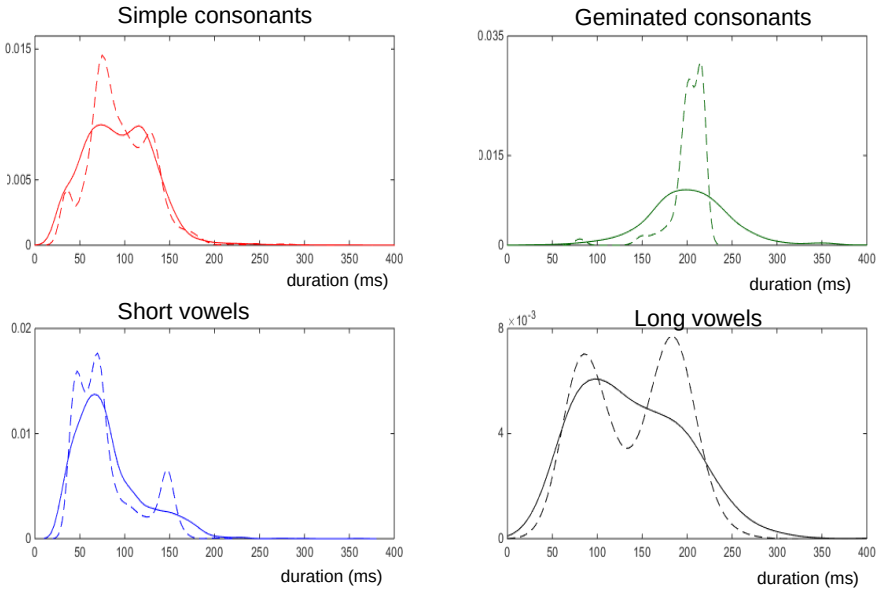

Fig. 3 Comparison of the probability distribution of reference (continuous line)and predicted (dotted line) duration for each class of phonemes

Subjective evaluation

In addition to the objective evaluation, listening tests were conducted to assess on one side the global quality of the uttered sounds, and on the other side the perception of phone durations. In both tests, 22 Arabic-native-speaking listeners, who are not speech specialists, were asked to rate the speech stimuli

Table 4 Comparison of RMSE, MAE and correlation between predicted duration and reference duration on the test set for the various modelling approaches

	Duration model	RMSE (ms)	MAE (ms)	Corr
All phones	HMM from HTS	30	20	0.83
	Class-specific DNN	**28**	**19**	**0.85**
	DNN-all-phone	32	22	0.80
	DNN-MERLIN	33	22	0.80
	MLP from [24]	41	28	0.66
All phones +pauses	HMM from HTS	40	24	**0.93**
	Class-specific DNN	**39**	**22**	**0.93**
	DNN-all-phone	42	25	0.92
	DNN from MERLIN	50	28	0.92
	MLP from [24]	63	37	0.87

using a score ranging from 1 (Very poor) to 5 (Very good), i.e. a Mean Opinion Score (MOS) test, and to compare pairs of stimuli, i.e. a preference test. Each participant has listened to 24 stimuli, 6 from each phone duration model, randomly selected amongst 96 stimuli (24 from each phone duration model). The subjective evaluation concerns the best duration model according to the objective evaluation above, plus two other duration modelling approaches and the reference duration signal:

- Natural, this is the reconstructed signal obtained after processing the original signal with the STRAIGHT vocoder (hence, corresponds to original durations).
- Class-specific-DNN, which is the signal synthesized after predicting the sound durations with the class-specific models, as proposed above.
- DNN-all-phone, which is the signal reconstructed using the durations predicted by the best DNN model trained over all the phonemes.
- HMM, which corresponds to the signal generated by HTS using the HMM-based duration model from HTS. This model was used as a baseline above.

It should be emphasized that the acoustic parameters (F0 and spectral parameters (MGC)) are generated from HMM models in HTS. The duration is generated externally from DNN duration models in the case of DNN-all-phone and Class-specific-DNN and from original duration in the case of Natural. In addition, it should be also mentioned, that all signals were generated with the HTS toolkit version 2.3 using the STRAIGHT vocoder. Also, it is worth noting that the subjective evaluation has been conducted following standard methodologies used for speech quality assessment, as detailed in [57,58].

Assessment of global quality

To assess this aspect, the 22 native Arabic listeners were asked to rate the overall quality of each stimuli through a MOS evaluation, and then through a preference test.

MOS Test Figure 4 shows that the class-specific-DNN duration model reaches the same MOS value as the natural duration reference. It is also interesting to note that the DNN-all-phone is ranked last, which means that to outperform HMM in duration modelling, the DNN should be used class-wise.

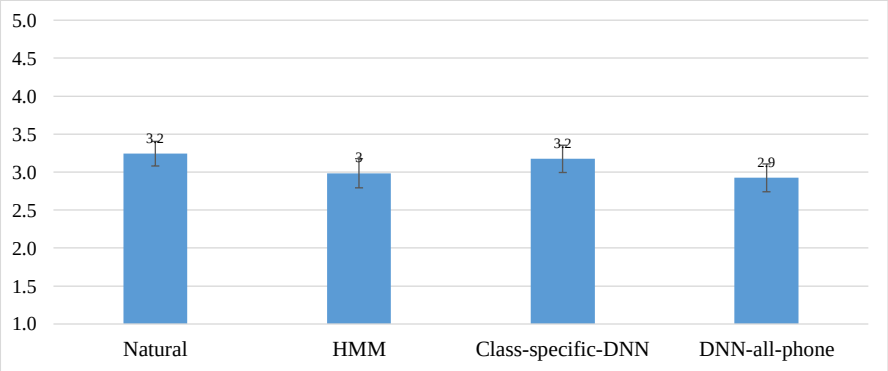

Fig. 4 MOS score results for global quality of generated speech

For a further analysis, the MOS values were finely investigated (cf. Figure 5), to show that this outstanding result for the class-specific-DNN model is due to the fact that it received a high MOS score (i.e. MOS $\geq$ 4) in more than 40% of cases. However, it is interesting to observe that HMM was the model which received the most of "Very good" scores (though the difference with Class-specific-DNN is slight).

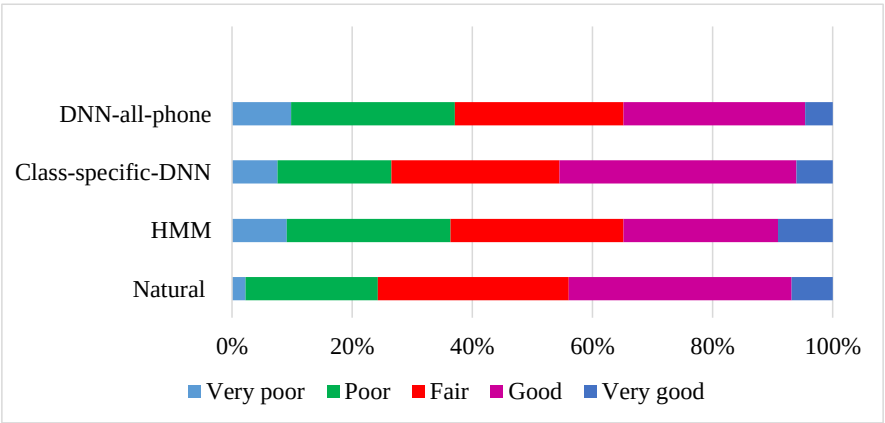

Fig. 5 Analysis of MOS scores for global quality of generated speech

Preference Test The MOS test results are confirmed by the preference test (cf. Figure 6), where listeners were asked to choose the sound having the best quality, for each pair of stimuli synthesized with different duration models. The main result is the clear preference of Class-specific-DNN above DNN-all-phone, and above HMM.

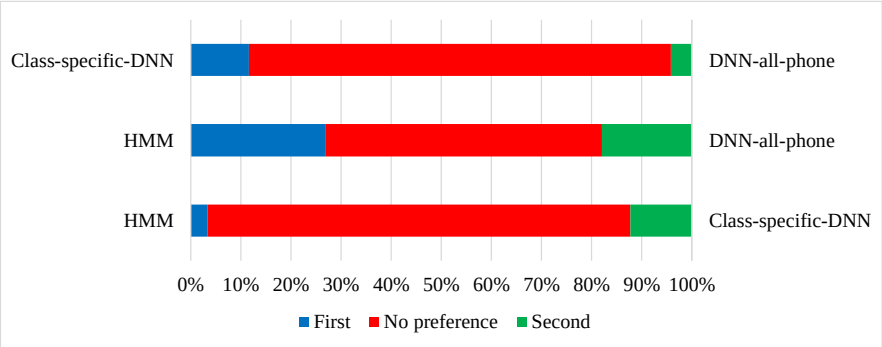

Fig. 6 Preference test results for global quality of generated speech

Assessment of duration perception

The second subjective test was conducted to assess the perception of phone duration. The 22 listeners were asked to rate the overall quality of the duration of the uttered phones through a MOS evaluation, and then to choose which stimuli is best with respect to phone duration in a preference test.

MOS test In this test, the listeners were asked to evaluate how they appreciate the duration of the pronunciation of the phones. The listeners had to choose one answer among five ranging from "phone durations are not respected" up to "phone durations are all respected". The MOS results are shown in Figure 7. It looks that HMM duration model was slightly preferred to the class-specific-DNN model. However, the difference is not significant when compared to the 95% confidence interval; and it is very relevant to note that both have a MOS result close to the one of the natural duration reference. In addition, Figure 8 shows a finer analysis of the given MOS scores. It is interesting to note that the class-specific-DNN model received as many high scores (MOS ≥ 4) as the HMM model, or the natural durations. Finally, the DNN-all-phone model was again less appreciated than all other models.

Preference test The participants were also asked to listen to pairs of stimuli, where for each pair, the same utterance is pronounced using a different duration model. They were asked to answer how they judge the duration of

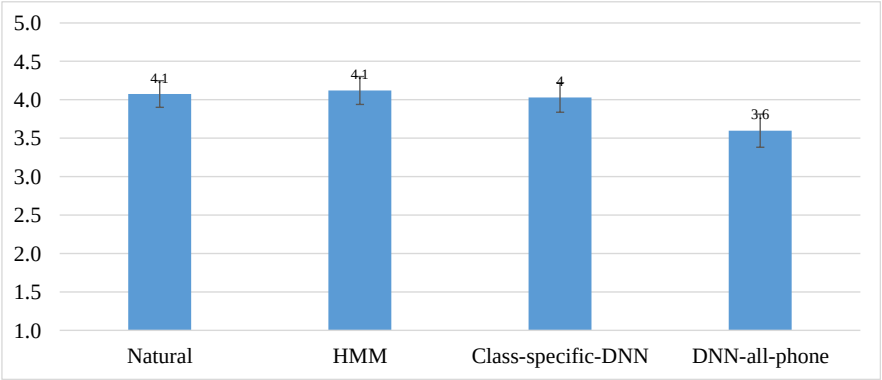

Fig. 7 MOS scores for duration perception

the pronunciation of the phones of the second stimuli in comparison to the first one. Figure 9 shows the results of the preference test. It appears that the HMM model and the class-specific-DNN model, are the most preferred, which matches with the overall quality assessment for duration pronunciation (cf. Figure 6).

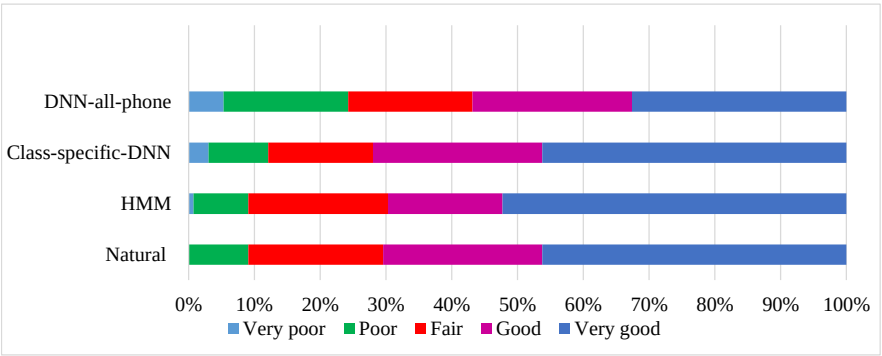

Fig. 8 Analysis of the MOS scores for duration perception

Interpretation and discussion

The results reported in the objective and the subjective evaluation can be analyzed further, from quantitative and qualitative viewpoints.

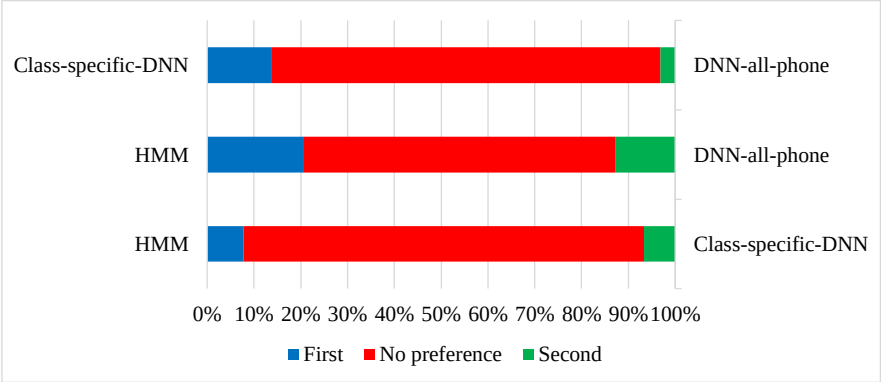

Fig. 9 Preference test results for duration perception

Quantitative analysis

The objective evaluation (cf. Table 3 and Table 4) shows that the proposed class-specific DNN model yields the least error measure, i.e. RMSE and MAE, and the highest correlation to the original duration values. This can be interpreted as a proof that each class of phonemes, i.e. short/long vowels, simple/geminated consonants has its own margin of duration. This is to recall a previous assumption, made by [59] who had stipulated that each phoneme has its own intrinsic duration, that can be stretched by an elasticity coefficient, whose value depends on the context. However, such an assumption has not been proved so far. Besides, a deeper cross-language investigation is necessary to reveal other contributory features affecting phone duration in other languages, so that the standard HTS features could be personalized for each target language.

Qualitative analysis

The analysis of MOS and preference test results (cf. Figure 4 to Figure 9) shows that speech reconstructed with the proposed specific-class DNN duration model is perceived as equal as that reconstructed with natural duration. Beyond this positive feedback, that was made by native Arabic speakers, a further analysis by listeners specialized in linguistics and phonetics should be carried out to judge the quality of each single type of phonemes. Actually, gemination is still a controverted phenomenon in Arabic phonetics, since phonetically, it is in the frontiers between a double consonant and a standalone unit. Also, vowel quantity is still a relative notion, since in Arabic, some long vowels might be pronounced as short ones, especially at the end of the word. Another point, which is not less important, consists in the necessity to study the effect such a phoneme-type-related modelling has on the other prosodic

parameters, such as F0 and spectrum, in correlation with the quality of parametric speech synthesis.

5 Conclusion

In this paper, a comprehensive investigation of duration modelling for Arabic sounds for TTS is described. A variety of DNN-based structures have been developed and evaluated. Taking into account some characteristics of Arabic, i.e. vowel quantity (short vs. long vowels) and gemination, the proposed DNN models were trained on different sets of phonemes, namely DNN-all-phone model that was trained on all the phonemes, whereas class-specific-DNN models were trained separately on each phoneme class, i.e. short/long vowels, simple/geminated consonants and pauses. The validation phase allowed selecting for each phoneme class the model that performed the best on the development set for evaluation on the test set. The main finding of the objective evaluation consists in observing that for each class of phonemes, the class-specific-DNN model performs better than the generic DNN model trained on all phonemes, i.e. DNN-all-phone model. This novel class-specific modelling approach was also compared to state-of-the-art models, i.e. baseline HMM from HTS toolkit and DNN from MERLIN toolkit. Again, the best accuracy values were obtained by the proposed class-specific DNN models, which results outperformed those of the baseline DNN from MERLIN in all cases, and were better or equal to those of HTS.

The subjective evaluation consisted in the assessment of the global quality of the generated sound and in the appreciation of the predicted sound durations. Both tests included a MOS and a preference rating. For performing these listening tests, the predicted duration obtained using the class-specific DNN model and the generic DNN model applied for all phonemes (DNN-all-phone) were processed with the HTS toolkit to generate the synthetic speech signals. In terms of global quality, the preference of listeners to the class-specific-DNN sounds confirms the objective evaluation results. Actually, class-specific-DNN was rated as high as natural, i.e. the original-duration-reconstructed sounds. This proves that the class-specific approach has succeeded to finely model the phone durations by fitting each DNN model to the corresponding class of phonemes, instead of the classical approach, i.e. applying one single DNN model for all phonemes. This proves also the benefit of handling specifically the various classes of Arabic sounds, through the addition of gemination and vowel quantity as special features for Arabic to the standard feature set of HTS, as done in previous studies, and / or through dedicated models as done in this current study for Arabic phone duration prediction.

6 Acknowledgments

This research work was conducted in the framework of PHC-Utique Program, financed by CMCU (Comité mixte de coopération universitaire), grant No.15G1405.

References

1. Klatt DH (1976) Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. The Journal of the Acoustical Society of America 59(5):1208-1221
2. Rubin P, Baer T, Mermelstein P (1981) An articulatory synthesizer for perceptual research. The Journal of the Acoustical Society of America 70(2):321-328
3. Moulines E, Charpentier F (1990) Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communication 9(5-6):453-467
4. Dutoit T, Leich H (1993) MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database. Speech Communication 13(3-4):435-440
5. Hunt AJ, Black AW (1996) Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Atlanta, GA, USA, pp 373–376
6. Yoshimura T, Tokuda K, Masuko T, Kobayashi T, Kitamura T (1999) Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In: Proc. European Conference on Speech Communication and Technology, Budapest, Hungary, pp 2347–2350
7. Imai S, Sumita K, Furuichi C (1983) Mel log spectrum approximation (MLSA) filter for speech synthesis. Electronics and Communications in Japan (Part I: Communications), 66(2):10-18
8. Zen H, Tokuda K, Black AW (2009) Statistical parametric speech synthesis. Speech Communication 51(11):1039-1064
9. Zen H, Senior A, Schuster M (2013) Statistical parametric speech synthesis using deep neural networks. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada, pp 7962–7966
10. Wang Y, Skerry-Ryan RJ, Stanton D, Wu Y, Weiss RJ, Jaitly N, Yang Z, Xiao Y, Chen Z, Bengio S, et al (2017) Tacotron: Towards end-to-end speech synthesis. In: Proc. Annual Conference of the International Speech Communication Association, Stockholm, Sweden, pp 4006–4010
11. Shen J, Pang R, Weiss RJ, Schuster M, Jaitly N, Yang Z, Chen Z, Zhang Y, Wang Y, Skerrv-Ryan Rj, et al (2018) Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: Proc. International Conference on Acoustics, Speech and Signal Processing, Calgary, Alberta, Canada, pp. 4779–4783
12. Mixdorff H(2002) An integrated approach to modeling German prosody, Doktor-Ingenieur habilitatus Dissertation, Technische Universitaet Dresden
13. Shinoda K, Watanabe T (1997) Acoustic modeling based on the MDL principle for speech recognition. In: Proc. European Conference on Speech Communication and Technology, Rhodes, Greece, pp 99–102
14. Abdelhamid O, Abdou SM, Rashwan M (2006) Improving Arabic HMM-based speech synthesis quality. In: Proc. International Conference on Spoken Language Processing, Pittsburgh, Pennsylvania, pp 1332–1335
15. Kawahara H (1997) Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. In: Proc. International Conference on Acoustics, Speech and Signal Processing, Munich, Germany, pp 1303–1306
16. Morise M, Yokomori F, Ozawa K (2016) WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE transactions on Information and Systems 99(7):1877-1884

17. Campbell WN (1993) Predicting segmental durations for accommodation within a syllable-level timing framework. In: Proc. European Conference on Speech Communication and Technology, Berlin, Germany, pp 1332–1335
18. Van Santen JP (1994) Assignment of segmentalduration in text-to-speech synthesis. Computer Speech and Language 8(2):95-128
19. Newman D (1984) The phonetics of Arabic. The Journal of the American Oriental Society 44:1-6
20. Abdelmalek R, Mnasri Z (2016) High quality Arabic text-to-speech synthesis using unit selection. In: Proc. IEEE International Multi-Conference on Systems, Signals, Signals & Devices, Leipzig, Germany, pp 1–5
21. Griffin DW, Lim JS (1988) Multiband excitation vocoder. IEEE Transactions on Acoustics, Speech, and Signal Processing 36(8):1223-1235
22. Boukadida F, Ellouze N (2005) Modélisation Statistique de la Durée des Voyelles en Parole Arabe. In: Proc. Science of Electronics, Telecommunications and Information Technology Conference, Tunisia, pp 1–4
23. Zaki A, Rajouani A, Najim M (2002) Un modèle prédictif de la durée segmentale pour la synthèse de la parole arabe à partir du texte. In: Proc. Journées d'Etudes sur la Parole, Nancy, France, pp 89–92
24. Mnasri Z, Boukadida F, Ellouze N (2009) Segmental duration modeling using non parametric statistical learning. International Review of Computer and Software 4(5):533-542
25. Houidhek A, Colotte V, Mnasri Z, Jouvet D, Zangar I (2017) Statistical modelling of speech units in HMM-based speech synthesis for Arabic. In: Proc. Language & Technology conference, Poznan, Poland, pp 1–6
26. Zangar I, Colotte V, Mnasri Z, Jouvet D, Houidhek A (2018) Duration modelling using DNN for Arabic speech synthesis. In: Proc. International Conference on Speech prosody, Poznan, Poland, pp 597–601
27. Yoshimura T, Tokuda K, Masuko T, Kobayashi T, Kitamura T (1998) Duration modeling for HMM-based speech synthesis. In: Proc. International Conference on Spoken Language Processing, Sydney, Australia, pp 29–32
28. Silén H, Helander E, Nurminen J, Gabbouj M (2010) Analysis of duration prediction accuracy in HMM-based speech synthesis. In: Proc. International Conference on Speech Prosody, Chicago, IL, USA, pp. 1–4
29. Zen H, Tokuda K, Masuko T, Kobayashi T, Kitamura T (2004) Hidden semi-Markov model based speech synthesis. In: Proc. International Conference on Spoken Language Processing, Jeju Island, Korea, pp 1393–1396
30. Pan S, Tao J, Wang Y (2011) A state duration generation algorithm considering global variance for HMM-based speech synthesis. In: Proc. Annual Summit and Conference, Asia Pacific Signal and Information Processing Association, Xi'an, China
31. Yijian W, Renhua W (2006) HMM-based Trainable Speech Synthesis for Chinese. Journal of Chinese Information Processing 20(4):75-81
32. Gao B, Qian Y, Wu Z, Soong FK (2008) Duration refinement by jointly optimizing state and longer unit likelihood. In: Proc. Annual Conference of the International Speech Communication Association, Brisbane, Australia, pp 2266–2269
33. Lu H, Wu YJ, Tokuda K, Dai LR, Wang RH (2009) Full covariance state duration modeling for HMM-based speech synthesis. In: Proc. International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, pp 4033–4036
34. Ishimatsu Y (2001) Investigation of state duration model based on gamma distribution for HMM-based speech synthesis. IEICE Technical Report, SP2001-81
35. Riley MD (1990) Tree-based modelling for speech synthesis. In: Proc. ESCA Workshop on Speech Synthesis, Autrans, France, pp 229–232
36. Rao KS, Yegnanarayana B (2007) Modeling durations of syllables using neural networks. Computer Speech and Language 21(2):282-295
37. Riedi M (1997) Modeling segmental duration with multivariate adaptive regression splines. In: Proc. European Conference on Speech Communication and Technology, Rhodes, Greece, pp 2627–2630
38. Lazaridis A, Honnet PE, Garner PN (2014) SVR vs MLP for Phone Duration Modelling in HMM-based Speech Synthesis, Technical Report No.EPFL-REPORT-198140

39. Ogbureke U, Cabral J, Berndsen J (2012) Explicit duration modelling in HMM-based speech synthesis using a hybrid hidden Markov model-Multilayer Perceptron. In: Proc. SAPA-SCALE Conference, Workshops on Statistical and Perceptual Audition Speech Communication with Adaptive Learning, Portland, OR, USA
40. Yu K, Mairesse F, Young S (2010) Word-level emphasis modelling in HMM-based speech synthesis. In: Proc. International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, pp 4238–4241
41. Chen B, Bian T, Yu K (2017) Discrete duration model for speech synthesis. In Proc. Annual Conference of the International Speech Communication Association, Stockholm, Sweden, pp 789–793
42. Chen B, Lai J, Yu K (2017) Comparison of Modeling Target in LSTM-RNN Duration Model. In: Proc. Annual Conference of the International Speech Communication Association, Stockholm, Sweden, pp 794–798
43. Wu Z, Watts O, King S (2016) MERLIN: An open source neural network speech synthesis system. In: Proc. ISCA Workshop on Speech Synthesis, Sunnyvale, USA, pp 202–207
44. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural computation 9(8):1735-1780
45. Henter GE, Ronanki S, Watts O, Wester M, Wu Z, King S (2016) Robust TTS duration modelling using DNNs. In: Proc. International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, pp 5130–5134
46. Fernandez, R, Rendel A, Ramabhadran B, Hoory R (2014) Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks. In: Proc. Annual Conference of the International Speech Communication Association, Singapore, pp 2268–2272
47. Zen H, Sak H (2015) Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In: Proc. International Conference on Acoustics, Speech and Signal Processing, Brisbane, Australia, pp 4470–4474
48. Moungsri D, Koriyama T, Kobayashi T (2017) Duration prediction using multiple Gaussian process experts for GPR-based speech synthesis. In: Proc. International Conference on Acoustics, Speech and Signal Processing, New Orleans, USA, pp 5495–5499
49. Halabi N, Wald M (2016) Phonetic inventory for an Arabic speech corpus. In: Proc. International Conference on Language Resources and Evaluation, Portoroz, Slovenia, pp 734–738
50. Rosen KM (2005) Analysis of speech segment duration with the lognormal distribution: A basis for unification and comparison. Journal of Phonetics 33(4):411-426
51. Sola J, Sevilla J (1997) Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Transactions on Nuclear Science 44(3):1464-1468
52. Halabi N (2016) Modern Standard Arabic Phonetics for Speech Synthesis. Dissertation, University of Southamtpon
53. HTS toolkit (2018) http://hts.sp.nitech.ac.jp. Accessed November 2018
54. MERLIN toolkit (2018) https://github.com/CSTR-Edinburgh/Merlin. Accessed November 2018
55. Arabic Speech Corpus (2020) http://en.arabicspeechcorpus.com/. Accessed August 2020
56. Wavenet (2020) https://deepmind.com/blog/article/wavenet-generative-model-raw-audio. Accessed August 2020
57. Thorpe LA, Shelton BR (1993) Subjective test methodology: MOS vs. DMOS in evaluation of speech coding algorithms. In: Proc. IEEE Workshop on Speech Coding for Telecommunications, pp 73–74
58. Dimolitsas S, Corcoran FL, Ravishankar C (1995) Dependence of opinion scores on listening sets used in degradation category rating assessments. IEEE Transactions on Speech and Audio Processing 3(5):421–424
59. Klatt DH, William EC (1975) Perception of segment duration in sentence contexts. In: Structure and process in speech perception, Springer, Berlin, Heidelberg, pp 69–89